SFU SIMON FRASER UNIVERSITY

Michael Smith Health Research BC

Natural Sciences and Engineering Research Council of Canada

Statistical considerations for consortia

and meta-analysis

Lloyd T. Elliott

Simon Fraser University

ICSDS 2025, Vancouver, June 2025

Outline

1. COVID-19 Host Genetics in Canada

2. Meta-analysis with overlapping studies

3. Variational inference for phylogenetics

Host Genetics

- COVID-19 RNA genome:
  - AUUAAAGGUUUAUACCUUCC ...
- Human DNA genome:
  - TAACCCTAACCCTAACCCTA ...

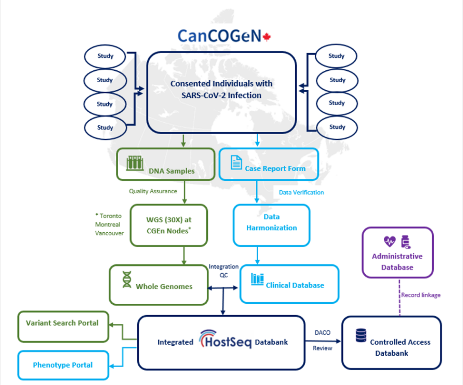
SNP: Single Nucleotide Polymorphism (can be coded as binary, or $\in \{0, 1, 2\}$)

How does human genetic variation modulate COVID-19 severity and susceptibility?
(Impact: therapy targets, understanding of disease)

HostSeq

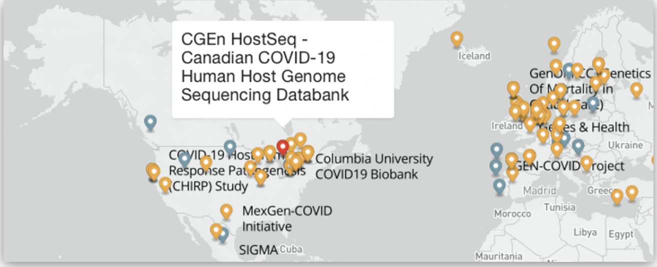
CGEn: Canada’s National Platform for Genome Sequencing and Analysis

- A project of projects
- 10,059 participants
- 15 study sites
- COVID-19 severity, phenotypes, covariates
- 33 approved DACOs (Summer 2025)



S. Yoo, E. Garg et al. HostSeq: A Canadian whole genome sequencing and clinical data resource. 2023. BMC Genomic Data. 24(26)

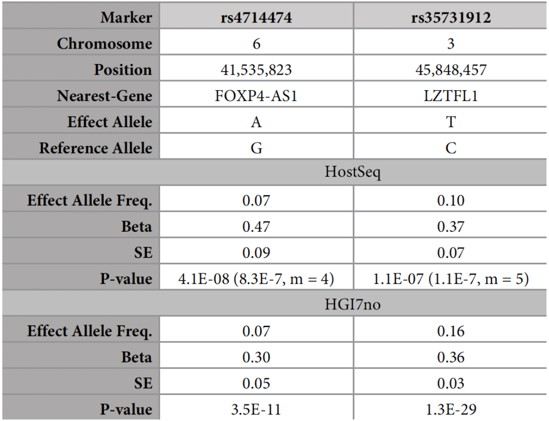
Host Genetics Initiative (HGI)



- Release 6 (June 2021)
  - 61 studies
  - 2,586,691 samples
- Release 7 (April 2022)
  - 81 studies (including HostSeq)
  - 2,942,817 samples

The COVID-19 Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. 2023. Nature. 621(7977). pE7-26

HostSeq GWAS



Marker	rs4714474	rs35731912
Chromosome	6	3
Position	41,535,823	45,848,457
Nearest-Gene	FOXP4-AS1	LZTFL1
Effect Allele	A	T
Reference Allele	G	C
HostSeq		
Effect Allele Freq.	0.07	0.10
Beta	0.47	0.37
SE	0.09	0.07
P-value	4.1E-08 (8.3E-7, m = 4)	1.1E-07 (1.1E-7, m = 5)
HGI7no		
Effect Allele Freq.	0.07	0.16
Beta	0.30	0.36
SE	0.05	0.03
P-value	3.5E-11	1.3E-29

E. Garg et al. Canadian COVID-19 host genetics cohort replicates known severity associations. 2024. PLOS Genetics. 20(3)

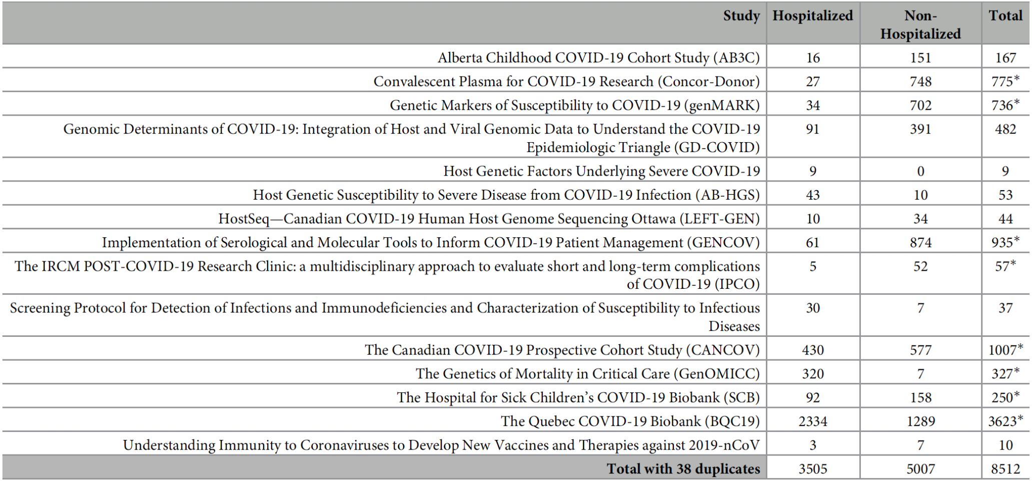
GWAS Results

- “We replicated two loci identified by the HGI for COVID-19 severity: the *LZTFL1/SLC6A20* locus on chromosome 3 and the *FOXP4* locus on chromosome 6 (the latter with a variant significant at $P<5E-8$).”

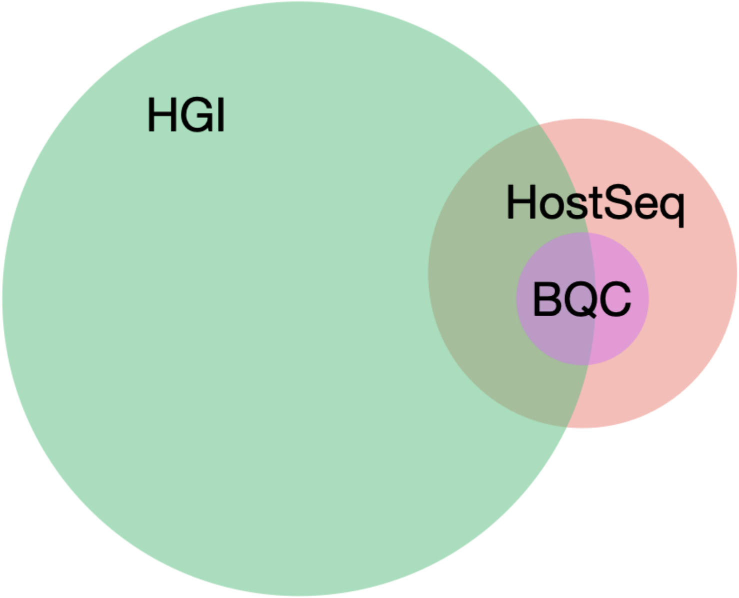
- “We found novel significant associations with *MRAS* and *WDR89* in gene-based analyses ...”

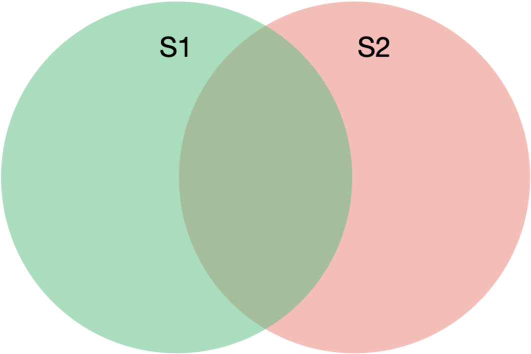
- “... and constructed a polygenic risk score that explained 1.01% of the variance in severe COVID-19.”

Overlapping Studies



Study	Hospitalized	Non-Hospitalized	Total
Alberta Childhood COVID-19 Cohort Study (AB3C)	16	151	167
Convalescent Plasma for COVID-19 Research (Concor-Donor)	27	748	775*
Genetic Markers of Susceptibility to COVID-19 (genMARK)	34	702	736*
Genomic Determinants of COVID-19: Integration of Host and Viral Genomic Data to Understand the COVID-19 Epidemiologic Triangle (GD-COVID)	91	391	482
Host Genetic Factors Underlying Severe COVID-19	9	0	9
Host Genetic Susceptibility to Severe Disease from COVID-19 Infection (AB-HGS)	43	10	53
HostSeq—Canadian COVID-19 Human Host Genome Sequencing Ottawa (LEFT-GEN)	10	34	44
Implementation of Serological and Molecular Tools to Inform COVID-19 Patient Management (GENCOV)	61	874	935*
The IRCM POST-COVID-19 Research Clinic: a multidisciplinary approach to evaluate short and long-term complications of COVID-19 (IPCO)	5	52	57*
Screening Protocol for Detection of Infections and Immunodeficiencies and Characterization of Susceptibility to Infectious Diseases	30	7	37
The Canadian COVID-19 Prospective Cohort Study (CANCOV)	430	577	1007*
The Genetics of Mortality in Critical Care (GenOMICC)	320	7	327*
The Hospital for Sick Children's COVID-19 Biobank (SCB)	92	158	250*
The Quebec COVID-19 Biobank (BQC19)	2334	1289	3623*
Understanding Immunity to Coronaviruses to Develop New Vaccines and Therapies against 2019-nCoV	3	7	10
Total with 38 duplicates	3505	5007	8512


HGI
HostSeq
BQC



- Summary statistics known for intersection: MetaSubtract. Nolte et al. 2017
- Known #cases/#control ratio: Dan-Yu Lin and Patrick F. Sullivan 2009
- Known size, unknown #case/#control ratio: ?

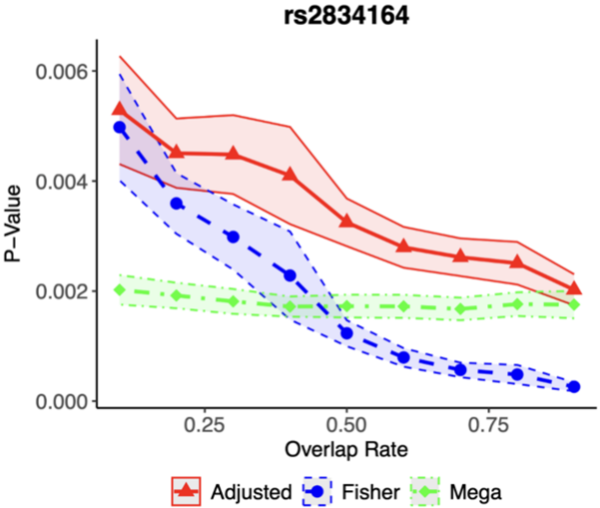
Derivation of an Adjusted Fisher method

- Let $\#S_1 = n_1$, $\#S_2 = n_2$, $\#S_1 \cap S_2 = n_{12}$
- Let Z_1, Z_2 be z-scores for the two studies, and p_1, p_2 their p-values

- Under the null, (Z_1, Z_2) is jointly normal with correlation ρ
- And $\widehat{\rho} = n_{12}/\sqrt{n_1 n_2}$ is an unbiased estimate of ρ (LeBlanc et al. 2018)
- Define $t_F = -2\log(p_1) - 2\log(p_2)$ (Fisher's method)
- Let $-2\log(p_1)$ and $-2\log(p_2)$ have correlation ρ^*
- By Ferrari et al. 2019, t_F is distributed as $\Gamma\left(\frac{m}{1+\rho^*}, 2(1+\rho^*)\right)$
- We use simulation to find an estimate of the mapping $\widehat{\rho} \to \rho^*$

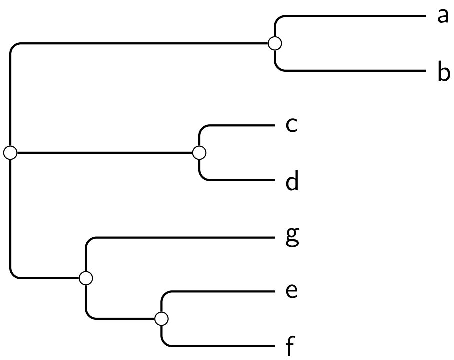
- Gives a p-value for t_F, yielding a *version of Fisher's method for overlapping samples*

Simulation based on HostSeq data



- Chose SNP at random with MAF > 0.1
- Phenotype, genotype from HostSeq
- Overlap simulated from 0.1 to 0.9 by choosing with 1,000 replicates containing 80% of the data

Phylogenetics



- Leaf nodes: Observed RNA sequences
- Interior nodes: Unobserved
- Goal: Infer tree structure
- Impact: Identify variants of concern, taxonomy, ...

Bayesian phylogenetics

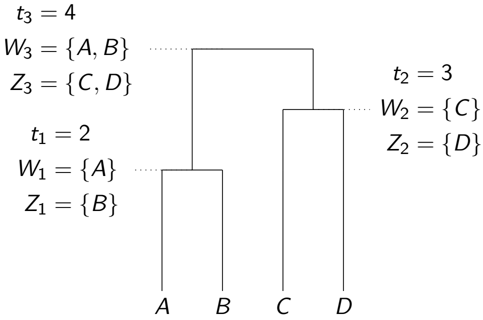
- Place a prior on the tree, develop proposals, perform MCMC (BEAST)

- Construct a variational family, perform *variational Bayes* (faster than MCMC?)
  - vB is an iterative inference algorithm that approximates the posterior using a family of functions (in contrast, MCMC approximates with a set of samples)

VIPR: Variational inference with products

over bipartitions

$$T = \begin{array}{c|cccc} & A & B & C & D \\ \hline A & * & \mathbf{2} & 8 & \mathbf{4} \\ B & * & * & 4.5 & 7 \\ C & * & * & * & \mathbf{3} \\ D & * & * & * & * \end{array}, \quad T = \begin{array}{cccc} A & B & C & D \\ \hline * & \mathbf{2} & 5 & 6 \\ * & * & \mathbf{4} & 7 \\ * & * & * & \mathbf{3} \\ * & * & * & * \end{array},$$



$t_3 = 4$

$W_3 = \{A, B\}$

$Z_3 = \{C, D\}$

$t_2 = 3$

$W_2 = \{C\}$

$Z_2 = \{D\}$

$t_1 = 2$

$W_1 = \{A\}$

$Z_1 = \{B\}$

Two possible example matrices T that could be drawn and result in the same phylogenetic tree after running single-linkage clustering. Entries of T that trigger coalescence are bolded

E. Sidrow, A. Bouchard-Côté and L.T. Elliott. Variational phylogenetic inference with products over bipartitions. 2025. ICML

The probability of a tree has a closed form ...

... in terms of the distribution of the distance matrix T

“If the random variables $t^{\{u,v\}}$ are mutually independent, and all $q_\phi^{\{u,v\}}$ are continuous in ϕ and t for all $\{u,v\}$ with $u,v \in$, and $Q_\phi^{\{u,v\}}$ is the survival function of $t^{\{u,v\}}$, then $q_\phi(\tau,)$ has the following form:

$$q_\phi(\tau,t) = \prod_{n=1}^{N-1} \left(\left(\sum_{\substack{w \in W_n \\ z \in Z_n}} \frac{q_\phi^{\{w,z\}}(t_n)}{Q_\phi^{\{w,z\}}(t_n)} \right) \prod_{\substack{w \in W_n \\ z \in Z_n}} Q_\phi^{\{w,z\}}(t_n) \right).”$$

Here W, Z is a bipartition induced by a coalescent, τ is the tree, t are the coalescent times (assuming Kingman). The inner sums and products have in total $\mathcal{O}(n^2)$ terms

Acknowledgements

Adjusted Fisher's method
Zikai Xu
Lin Zhang

VIPR
Evan Sidrow
Alexandre Bouchard-Côté

HostSeq collaboration
Elika Garg
Olga Vishnyakova
Paola Arguello Pascualli
Steve Jones
Lisa Strug
Jennifer Brookes
Shelley Bull
France Gangnon
Celia Greenwood
Rayjean Hung
Jerry Lawless
Andrew Patterson
Lei Sun
Study PIs
Sub-Committees
Study participants

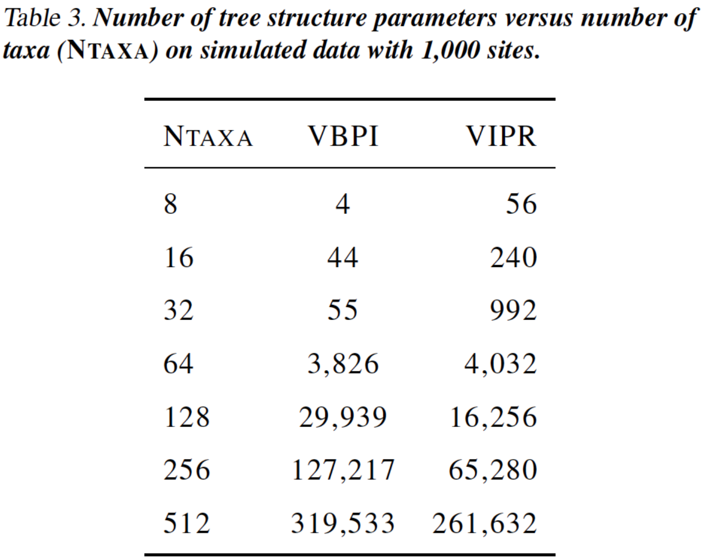


Table 3. ***Number of tree structure parameters versus number of taxa (NTAXA) on simulated data with 1,000 sites.***

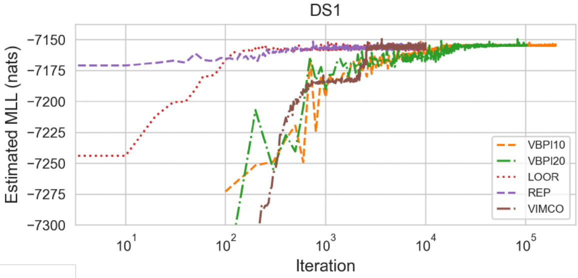
NTAXA	VBPI	VIPR
8	4	56
16	44	240
32	55	992
64	3,826	4,032
128	29,939	16,256
256	127,217	65,280
512	319,533	261,632

Thank You!

Table 6. ***Gap between gold standard and estimated marginal log-likelihoods for variational inference methods (in nats).*** *Marginal log-likelihoods for VI methods were estimated using importance sampling with 1,000 random samples from each variational distribution. Values indicate differences between gold standard MLLs and each method's MLLs. Gold standard MLLs (indicated by the* BEAST *column) are derived from 10 independent chains of the stepping-stone algorithm in BEAST. Datasets (*Data *column) DS1 to DS11 are from Lakner et al. (2008). Dataset COV is the COVID-19 dataset obtained from GISAID. VI methods are specified by columns: Variational Bayesian Phylogenetic Inference with K-sample ELBO, $K = 10$ (*VBPI10*; Zhang and Matsen IV 2024); Variational Bayesian Phylogenetic Inference with K-sample ELBO, $K = 20$ (*VBPI20*; Zhang and Matsen IV 2024); VIPR using the leave-one-out REINFORCE estimator (*LOOR*); VIPR using the reparameterization trick (*REP*); VIPR using the Variational Inference for Monte Carlo Objectives estimator with $K = 10$ (*VIMCO*). Standard errors were estimated using 100 bootstrapped samples and are shown in parentheses.*

Data	(N, M)	BEAST	VBPI10	VBPI20	LOOR	REP	VIMCO
DS1	(27, 1949)	−7154.26(0.19)	−0.53(0.09)	**0.36**(**0.13**)	−2.29(0.15)	−1.83(0.21)	−0.95(0.46)
DS2	(29, 2520)	−26566.42(0.26)	**0.16**(**0.24**)	0.01(0.20)	−0.76(0.14)	−0.14(0.43)	−0.37(0.29)
DS3	(36, 1812)	−33787.62(0.36)	−0.44(0.12)	−**0.38**(**0.13**)	−3.66(0.53)	−1.91(0.99)	−2.63(0.50)
DS4	(41, 1137)	−13506.05(0.32)	0.03(0.53)	**0.46**(**0.43**)	−2.48(0.43)	−0.47(1.21)	−1.73(0.23)
DS5	(50, 378)	−8271.26(0.39)	−1.70(0.35)	−5.69(0.48)	−0.29(1.82)	−4.01(0.28)	**0.94**(**2.08**)
DS6	(50, 1133)	−6745.31(0.55)	−0.76(0.20)	−**0.32**(**0.35**)	−3.96(0.34)	−3.26(0.60)	−2.72(0.37)
DS7	(59, 1824)	−37323.88(0.66)	**0.27**(**0.26**)	−0.24(0.17)	−2.73(0.30)	−2.82(0.31)	−10.42(0.70)
DS8	(64, 1008)	−8650.20(0.77)	−0.82(0.27)	**0.47**(**0.64**)	−3.28(0.99)	−4.95(0.47)	−2.88(0.60)
DS9	(67, 955)	−4072.66(0.53)	−5.32(0.31)	−4.12(0.46)	−**3.12**(**1.21**)	−5.79(0.74)	−7.60(0.44)
DS10	(67, 1098)	−10102.65(0.65)	−**0.88**(**0.20**)	−1.44(0.22)	−5.38(0.42)	−3.98(1.14)	−6.82(0.49)
DS11	(71, 1082)	−6272.57(0.68)	−18.79(0.41)	−16.28(0.46)	−**6.79**(**0.89**)	−7.31(0.71)	−9.62(1.46)
COV	(72, 3101)	−7861.61(0.74)	−39.08(0.58)	−**33.26**(**0.76**)	−611.84(1.80)	−374.62(0.48)	−214.25(0.42)

DS1



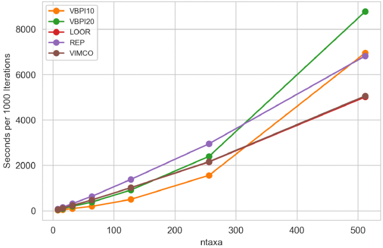
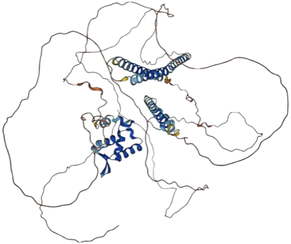


Figure 3. ***Seconds per 1,000 iterations vs. number of taxa.*** *Each VI method was run for 1,000 iterations or 5 minutes (whichever took less) on simulated datasets.*

Gene card: *FOXP4*

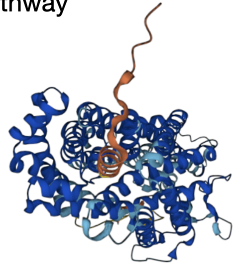
- Forkhead Box P4. A protein coding gene (coding Forkhead Box Protein P4).
- Located on 6p21.1, 56k bases long
- Gene transcription regulator, with some lung-specific regulation. May be involved in repressing some lung-specific expression. "... involved in the upkeep of healthy lung tissue ..."



genecards.org, AlphaFold and HGI

Gene card: *SLC6A20*

- Solute Carrier Family 6 Member 20. A protein coding gene (coding protein of same name)
- Located on 3p21.31, 41k bases long
- Transports small molecules across the cell membrane (prolines). Identified as a viral entry pathway



genecards.org, AlphaFold and HGI